UCRL-PROC-203118

Effect of Colorspace Transformation, the Illuminance Component, and Color Modeling on Skin Detection

S. Jayaram, S. Schmugge, M. C. Shin, L. V. Tsap

March 24, 2004

IEEE Computer Society Conference on Computer Vision and Pattern Recognition
Washington, DC, United States
June 27, 2004 through July 2, 2004

Disclaimer

Effect of Colorspace Transformation, the Illuminance Component, and Color Modeling on Skin Detection

Sriram Jayaram, Stephen Schmugge, Min C. Shin
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC 28223
sjayaram, sjschmug, or mcshin@uncc.edu

Leonid V. Tsap
Electronics Engineering Department
Lawrence Livermore National Laboratory
Livermore, CA 94551
tsap@llnl.gov

Abstract

Skin detection is an important preliminary process in human motion analysis. It is commonly performed in three steps: transforming the pixel color to a non-RGB colorspace, dropping the illumination component of skin color, and classifying by modeling the skin color distribution. In this paper, we evaluate the effect of these three steps on the skin detection performance. The importance of this study is a new comprehensive colorspace and color modeling testing methodology that would allow for making the best choices for skin detection. Combinations of nine colorspaces, the presence of the absence of the illuminance component, and the two color modeling approaches are compared. The performance is measured by using a receiver operating characteristic (ROC) curve on a large dataset of 805 images with manual ground truth. The results reveal that (1) the absence of the illuminance component decreases performance, (2) skin color modeling has a greater impact than colorspace transformation, and (3) colorspace transformations can improve performance in certain instances. We found that the best performance was obtained by transforming the pixel color to the SCT, HSI, or CIELAB colorspaces, keeping the illuminance component, and modeling the color with the histogram approach. [1]

1. Introduction

Skin detection is an important preliminary process in human motion analysis and face detection techniques. Many skin detection applications are used in environments with a considerable variation in the skin tones, the amount of illumination, and the type of illumination. Frequently, skin detection is performed in three steps of (1) transforming the color of pixel to another colorspace, (2) dropping the illuminance component of the colorspace and using only two color components in the classification process, and (3) classifying by modeling the distribution of skin color. Such steps are assumed to provide a robust performance under different lighting conditions and skin tones. However, the claimed benefit of the robust performance has not been rigorously examined.

In this paper, to enable the users of skin detectors to choose the best approach, we have attempted to examine the effect of these three steps by using a receiver operating characteristic (ROC) curve. First, we examined if better performance can be achieved by transforming the color of pixel to another colorspace, or dropping the illuminance component. Second, we examined if skin color modeling approaches make a difference in performance. Lastly, we ranked the performance of the combination of the nine colorspaces, the presence or the absence of the illuminance component, and two color modeling approaches. The performance is measured by using a ROC curve on a large dataset of 805 images including skin pixels taken from different skin tones under different lighting conditions. The advancements of this work over our previous work (not cited to preserve anonymity) are that (1) we use the train-test paradigm in conjunction with a 10-fold cross validation, and (2) we examine the combinations of the three steps of skin detection using the ROC analysis. We believe that the conclusions of this work will be helpful for the users of skin detectors in selecting the best method.

2. Previous Work

2.1. Comparative Studies

Terrillon et al. [11] compared the performance of nine chrominance spaces in the context of face detection. The colorspaces were evaluated using two metrics. First was the Mean Square Deviation (MSDN) which was computed from the normalized histogram constructed for each colorspace. Second metric was the degree of overlap between the skin and non skin distributions referred called HIN. The paper concluded that the discrimination between the skin

[1]The work was funded in part by Sun Microsystems Grant EDUD-7824-030480-US. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract number W-7405-Eng-48. UCRL-JC-~~149946~~

class and the non-skin class is the highest in the normalized spaces. Albiol et al. [1] claimed that for every colorspace there is an optimum skin detector with comparable performance. A skin detector can be found with the same performance regardless of the colorspace chosen, provided that there is an invertible transformation between the compared colorspaces. Three colorspaces of RGB, YCbCr, and HSV were compared. The paper concluded that (1) the operating characteristics of the three dimensional colorspaces were the same, and (2) the performance of the CbCr colorspace was lower since the transformation from any 3 dimensional colorspace to bi-dimensional colorspace is not invertible. Zarit et al. [13] summarized two different approaches of pixel classification in five different colorspaces. Two classification approaches were the look up table (LT) and the Bayesian decision theory. The five colorspaces of HSV, NRGB, CIELAB, Fleck-HS and YCbCr were compared. The HSV and Fleck-HS had the best performance in all tests. The CIELAB and YCbCr had the worst results. The Bayesian results of the colorspaces were similar. The LT method preformed best when used with the HSV or the Fleck-HS.

2.2. Skin classification with illumination invariance studies

Yang et al. [12] investigated the human skin color distribution using the images of 48 human faces. The study verified that the color differences among people lie in the intensity rather than the color itself. Thus the color difference can be minimized by color normalization. The study also concluded that the skin color clustered as Gaussian-like distribution by referring to the plots of the RGB values of skin colors. Furthermore, it was observed that the variance of skin color was lower in the normalized colorspace than the original RGB and that skin color of an individual can be characterized by multivariate normal distribution under certain lighting. The paper proposed a method to detect human faces by modeling the skin color using CIE LUV with luminance component L discarded. A set of 500 face images from different ethnic background, different orientation, and complex background were used to build a color histogram. The paper noted that the skin color distribution could be modeled by the Gaussian distribution. The results showed that the human faces can be robustly detected irrespective of orientation, size and viewpoint. Hsu et al. [5] proposed a face detection algorithm in color images under varying illumination and complex background. To accommodate varying illumination, the lighting was compensated by scaling the average gray value of pixels with top 5% of the luma linearly to 255. The YCbCr colorspace was used. The algorithm was tested on images containing faces with a complex background with a variation in color, position, scale, orientation, 3D pose, facial expression, and illumination.

3. Experimental Methods

3.1. Colorspace Transformations

In our work, we assume that color is represented in three dimensions, and colorspace transformation is a process of converting the pixel color in the RGB to another colorspace. The images in our dataset were captured in the RGB colorspace. The performance of nine colorspaces (CIELAB, CIEXYZ, HSI, NRGB, RGB, SCT [10], YIQ, YUV, YCbCr) is evaluated. The transformation to a non-RGB colorspace is performed through a simple or nonlinear transformation of the RGB [14]. All color components are normalized to the range of [0, 255] and quantized to 256 levels.

3.2. Dropping Illuminance Component

It is often assumed that some robustness in the skin detection performance may be achieved by using color without the illuminance component. We define the **3D color** as the color with all three color components and the **2D color** as the color with two chrominance components (without the illuminance component.) The illuminance components of the colorspaces are B for RGB and NRGB, L for CIELAB, Y for CIEXYZ, I for HSI, L for SCT, and Y for YIQ, YCbCr, and YUV.

3.3. Modeling of Skin Color

We used a classifier incorporating the Bayesian decision theory to classify a pixel color into skin class (ω_s) or non-skin class (ω_{ns}). For each pixel, a feature ($\mathbf{x}$) is created by using all three color components after colorspace transformation for the 3D color or by using two chrominance components after colorspace transformation for the 2D color. The posterior probability of ω_s is computed as

$$p(\omega_s|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_s)p(\omega_s)}{p(\mathbf{x})}$$

where $p(\mathbf{x}|\omega_s)$ is class condition probability, $p(\omega_s)$ is the priori probability, and $p(\mathbf{x})$ is the evidence factor. Each feature vector ($\mathbf{x}$) is classified as ω_s if $p(\omega_s|\mathbf{x}) > T_{skin}$, or ω_{ns} otherwise. We examined two methods of modeling the skin color : a normal density and histogram approach.

3.3.1 Normal Density

Using the normal density approach, the class conditional probability of a class is determined using the multi-variate normal density equation

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)]$$

where $\mathbf{x}$ is a d-component column vector, μ is the d-component mean vector, Σ is dxd covariance matrix [4].

3.3.2 Histogram approach

In the histogram approach [6], the probability is modeled with a histogram for each class. The histograms are quantized into bins per channel. The class conditional probability $p(\mathbf{x}|\omega_s)$ is computed as the ratio of each bin value ($c[\mathbf{x}]$) to the sum of all bins. During the training of 10-fold cross validation, the best histogram size (number of bins per color channel) is found for each combination of colorspace and the presence (or absence) of the illumination component. The histogram sizes of 8, 16, 32, 64, 128, and 256 were examined. More discussion on the effect of histogram size is presented in Section 4.5.

3.4. Dataset

Figure 1: Top row : skin images from AR & UOPB dataset. Bottom row : non-skin images from UW dataset.

A dataset of 805 images with 6.5 million pixels was used to compute the performance. The dataset was composed of 510 images (3.2 million pixels) with skin pixels and 298 images (3.3 million pixels) without skin pixels. A sample images are shown in Figure 1. The images with skin pixels were collected from the AR face dataset [9] and the UOPB dataset [8]. We did not include the non-skin pixels from AR and UOPB dataset, because those images were taken with white background. Instead, we collected the images without skin from the University of Washington content-based image retrieval database [2]. The AR face dataset includes more than 4000 frontal color images of 70 men and 56 women with different facial expression, and four illumination levels. The images were taken at two different times in a span of 14 days. The UOPB dataset includes 2,112 frontal images of 111 different people in 16 camera calibration and illumination condition of horizon, incandescent, fluorescent and daylight.

3.5. Ground Truth

Figure 2: A sample of ground truth (left) and corresponding color image (right). Skin pixels are colored in black. Difficult and tedious regions to mark (‘don’t care’) are colored in gray (shown in darker gray.) The background in the skin images is also marked in gray (shown in lighter gray) indicating that those pixels did not participate in the evaluation. Note that the non-skin pixels were collected in non-skin images from [2]

The ground truth (GT) is defined at the pixel-level. The three labels method similar to one used for the edge detection evaluation study [3] is used to label pixels as skin (black), non-skin (white), or don’t-care (gray) as shown in Figure 2. ‘Don’t care’ label is assigned to pixels that are too ambiguous or tedious to label as either skin or non-skin. It is difficult to accurately label the boundary pixels between skin and non-skin regions. So, we label the pixels with the width of 5 pixels along the boundary as ‘don’t care’. Since the background of the skin images were mostly white, the pixels of the background are not used as non-skin data. The background pixels of skin images are labeled in gray (see Figure 2) indicating that they were not used for the experiments.

3.6. Performance Metrics

There are 36 combinations from nine colorspaces, the absence or the presence of the illuminance component (the 2D and 3D color), two color modeling approaches. For each combination, we trained and tested using a 10-fold cross validation. The dataset of 805 images were randomly divided into 10 folds. The average of test performance of 10 folds is used to evaluate the performance.

The class conditional probability is computed using the training set. We computed performances using 50 different

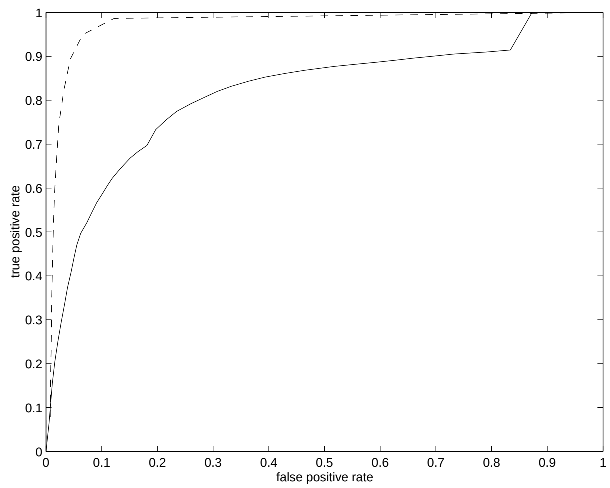

Figure 3: The ROC curves of the best performing combination (SCT 3D with histogram) is shown in dash line and the worst performing (RGB 2D with normal density) combination is shown in solid line.

values of T_{skin} (the threshold used for classification) between [0, 1]. For each value of T_{skin}, the pixels in testing set were classified as ω_s or ω_{ns}. Then, the classified pixels were compared against the ground truth to determine the total number of true positive (TP) and false positive (FP). The TP and FP were normalized by dividing by the total number skin pixels and non-skin. The coordinate of [normalized FP, normalized TP] were used to describe the performance of skin detection for a given T_{skin}. Then, the ROC curve was formed by taking the locations with the highest normalized TP for a given normalized FP. The area under curve (AUC) is used to compare the performance. The ROC curves of the best and the worst performing combinations are shown in Figure 3. Refer to Table 1 for the rankings.

4. Results

In this section, we attempt to answer four questions by using the testing performances of 36 combinations. Each question is addressed in the subsections.

4.1. Does color transformation help?

We examined if a color transformation improves the performance (refer to Table 1.) We found that the colorspace transformation does help on two cases: when the 2D color was modeled using the normal density approach (8 out of 8 improved) and when the 3D color was modeled using the histogram approach (6 out of 8 improved). Only about half of transformations helped on other two cases: when the 3D color was modeled using the normal density approach (3 out of 8 improved) and when the 2D color was modeled using the histogram approach (4 out of 8 improved). Overall, the improvement in performance due to the colorspace transformation was present, but not consistent.

Table 1: The testing performance of 36 combinations are shown in AUC (area under curve). Each AUC is an average from ten testings of a 10-fold cross validation. Higher AUC value indicates better performance. The ranking is shown in the parenthesis. For the histogram approach, the performance is based on the trained histogram size.

	normal		histogram	
colorspace	3D	2D	3D	2D
CIELAB	0.907 (29)	0.908 (23)	0.977 (3)	0.960 (11)
CIEXYZ	0.908 (28)	0.914 (21)	0.959 (12)	0.935 (15)
HSI	0.929 (16)	0.934 (19)	0.980 (2)	0.973 (4)
NRGB	0.912 (22)	0.917 (20)	0.955 (13)	0.963 (9)
RGB	0.908 (24)	0.826 (36)	0.960 (10)	0.949 (14)
SCT	0.931 (18)	0.932 (17)	0.982 (1)	0.968 (5)
YCbCr	0.908 (26)	0.861 (33)	0.964 (7)	0.886 (31)
YIQ	0.908 (27)	0.861 (34)	0.966 (6)	0.887 (30)
YUV	0.908 (25)	0.841 (35)	0.965 (8)	0.886 (32)

4.2. Does illuminance component dropping help?

We examined if the performance improved when the illumination component of color was not used. The pair-T test results are shown in the upper half of Table 2. With both modeling approaches, the performance with the presence of the illuminance component (the 3D color) was significantly better than the 2D color at the 95% confidence interval. The performance difference between the 3D and 2D color was larger and more significant with the histogram-based approach than the normal density approach. We conclude that the absence of the illuminance component did not help. Interestingly, the rankings of the HSI and SCT did not change much between the 3D color and the 2D color (Table 1.) Thus for those colorspaces, we suspect that the illuminance component does not add much information to the separation of skin from non-skin. However, when the YCbCr, YIQ, and YUV are used with the histogram modeling, the ranking degraded greatly from the 3D color to 2D color. In fact, the degradation of those colorspaces was largest of all colorspaces.

4.3. Does the skin color modeling make any difference?

We also examined if different skin color modeling approaches make a difference on the detection performance (refer to the lower half of Table 2.) We found that the histogram-based approach was significantly better than the normal density approach at the 95% confidence interval. In fact, the difference due to different skin color modelings

Table 2: Paired Samples T-Test on performance of two classifiers.

compared	paired differences					t	df	sig
				95% confidence interval of the difference				
	mean	std. dev	std. error mean	lower	upper			
Normal3D - Normal2D	.02502	.035595	.011865	-.00234	.05239	2.109	8	.068
Histogram3D - Histogram2D	.03122	.037598	.012533	.00232	.06012	2.491	8	.037
3D - 2D	.0281	.03566	.00841	.0104	.0459	3.346	17	.004
Normal3D - Histogram3D	-.05079	.009814	.003271	-.05833	-.04324	-15.525	8	.000
Normal2D - Histogram2D	-.04459	.031731	.010577	-.06898	-.02020	-4.215	8	.003
Normal - Histogram	-.0477	.02301	.00542	-.0591	-.0362	-8.7494	17	.000

was twice the difference due to the presence and the absence of the illuminance component. We conclude that the choice of skin color modeling does make a significant difference. Also, its effect on the performance is greater than the effect due to the presence or the absence of the illuminance component.

4.4. Which is the best combination of colorspace, color dimension, and color modeling?

Thirty-six combinations were ranked using the testing performance. The rankings are shown in parenthesis in Table 1. The best performance was achieved by the combination of the SCT colorspace, with the presence of the illuminance component, and modeling the color using the histogram approach. The performance of the HSI (the second) and the CIELAB (the third) was also very good as well. The worst performance was achieved with the RGB with the absence of the illuminance component using the normal density approach. The ROC curves of the best and the worst performing combinations are shown in Figure 3. Overall, we conclude that a good skin detection performance was achieved by using the SCT, HSI or CIELAB colorspaces with the presence of the illuminance component, and modeling the color using the histogram approach.

4.5. Choice of Histogram Size

We have examined the effect of the histogram size (defined in Section 3.3.2) on the performance. The testing performance by using the histogram sizes of 8, 16, 32, 64, 128, and 256 are shown on Table 3. The best performing histogram size is highlighted in bold for each colorspace and color dimension (the 2D or 3D color.) First, note that the performance degrades significantly when a small histogram size is used. The amount of degradation was larger with the 2D color than the 3D color. Interestingly, some colorspaces (such as SCT) degraded much less than others (such as YCbCr, YIQ, YUV). It is beyond the scope of this paper to examine the reasons behind such findings, but it is included in our future research. Second, in general, the best histogram sizes were smaller with the 3D color than the 2D color. On average, the best histogram size was 64 for the 3D color and 128 for the 2D color.

5. Discussion

Jones and Rehg [6] found that their histogram-based skin detector performed better than the detector based on a mixture of Gaussian. Although we did not explicitly compare the mixture of Gaussian, we did find that their method performed better than the Gaussian-based method. However, we noted that the authors could have further improved their detector by choosing the best colorspace. We observed that their choice of colorspace and dimension (the RGB with the illuminance component) with the histogram approach was ranked to be the 11th out of 18 (Table 1.) This clearly demonstrates that our work can benefit the skin detection methods by assisting in the selection of all three steps of skin detection.

Powell and Murphy [10] demonstrated that the color segmentation using the SCT colorspace transformation without the illuminance component is better than the colorspaces of the HSI and RGB. In this work, we found that the SCT colorspace has performed well for the task of skin detection (refer to Table 1.) However, without the illuminance component, the HSI performed slightly better than the SCT. Interestingly, *with* the illuminance component, the SCT was slightly better than the HSI.

In our previous work (not cited to preserve anonymity), we have examined the effect of colorspace transformation with respect to the separability of skin color from non-skin color. In both previous work and this work, we found that

Table 3: The testing performance of histogram sizes of 8, 16, 32, 64, 128, and 256. The best performing histogram size is highlighted in bold for each colorspace and color dimension (the 2D or 3D color.)

colorspace	the 2D color						the 3D color					
	8	16	32	64	128	256	8	16	32	64	128	256
CIELAB	0.656	0.874	0.936	0.955	**0.960**	0.959	0.836	0.935	0.968	**0.977**	0.974	0.961
CIEXYZ	0.845	0.899	0.922	0.925	0.930	**0.935**	0.830	0.899	0.941	0.955	**0.962**	0.956
HSI	0.937	0.953	0.971	**0.973**	0.971	0.965	0.953	0.970	**0.980**	0.976	0.968	0.948
NRGB	0.917	0.919	0.949	0.950	**0.963**	0.961	0.903	0.943	0.950	0.948	0.956	**0.957**
RGB	0.840	0.899	0.930	0.942	**0.949**	0.934	0.907	0.949	0.958	**0.962**	0.959	0.919
SCT	0.932	0.947	0.963	0.967	**0.968**	0.963	0.945	0.979	**0.982**	0.977	0.967	0.938
YCbCr	0.635	0.834	0.870	0.880	0.884	**0.886**	0.860	0.944	**0.964**	**0.964**	0.963	0.955
YIQ	0.621	0.800	0.865	0.878	0.885	**0.887**	0.860	0.939	**0.965**	0.963	0.963	0.947
YUV	0.635	0.834	0.870	0.880	0.884	**0.886**	0.860	0.944	0.964	**0.965**	**0.965**	0.955
average	0.780	0.884	0.920	0.928	**0.933**	0.931	0.884	0.945	0.964	**0.965**	0.964	0.949

the performance decreases significantly from the presence of the illuminance component to the absence. However, we found that the improvement due to colorspace transformation was rare (according to the separability study) and existing but not consistent in this study. Differences between the two show that separability and classification do not always correlate. Also, this work involves actually testing on the separate validation dataset where as the previous work compares the separability of one whole dataset.

6. Conclusions

Skin detection is an important preliminary process in human motion analysis and face detection. Our work is novel in that we provide a comprehensive testing methodology of those three steps that would allow for selecting the best skin detection approach. Our conclusions are drawn from a large dataset with manually defined ground truth using a ROC curve. First, our extensive analysis reveals that the colorspace transformation does improve the performance, but not consistently. The improvement varied among the combination of three steps. Second, the performance is significantly better with the presence of the illuminance component. Third, the skin color modeling using the histogram approach was significantly better than the Gaussian approach. In fact, the performance difference between the modeling approach was twice the difference due to the presence or the absence of the illuminance component. Fourth, we found that the the best skin detection performance was obtained by transforming the pixel color to the SCT, HSI, or CIELAB colorspace and modeling the color with the illuminance component using the histogram-based approach. By evaluating all three steps thoroughly, we have provided a tool for selecting the best approach for skin detection.

References

[1] Alberto Albiol, Luis Toress, Edward Delp, "Optimal color spaces for skin detection," *ICIP*, 2001,

[2] A. Berman and L. G. Shapiro, "A Flexible Image Database System for Content Based Retrieval," CVIU, Vol. 75, Nos. 1-2, pp. 175-195, 1999.

[3] K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge Detector Evaluation Using Empirical ROC Curves," CVIU. 84 (1), October 2001.

[4] Richard o. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2001, pp 57-100.

[5] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, Anil K. Jain, "Face Detection in Color Images", TPAMI, Vol.24, No.5, pp 696-706, 2002.

[6] Michael J. Jones, James M. Rehg, "Statistical Color Models with Application to Skin Detection", IJCV, Vol. 46, No. 1, January 2002, pp. 81-96.

[7] Helena Chmura Kraemer, Evaluating Medical Tests: Objectives and Quantitative Guidelines, pp 55-100, 1992.

[8] E. Marszalec, B. Martinkauppi, M. Soriano, and M. Pietikinen, "A physics-based face database for color research," Journal of Electronic Imaging, Vol. 9 No. 1 pp. 32-38.

[9] A. M. Martinez and R. Benavente, "The AR Face Database", CVC Technical Report #24, June 1998.

[10] M. W. Powell and R. Murphy. "Position estimation of micro-rovers using a spherical coordinate transform color segmenter." In *IEEE Workshop on Photometric Modeling for Computer Vision and Graphics*, pages 21–27, Fort Collins, CO, June 1999.

[11] Jean-Christophe Terrillon and Shigeru Akamatsu, "Comparative Performance of Different Chrominance Spaces for Color Segmentation and Detection of Human Faces in Complex Scene Images," *International Conf on Face and Gesture Recognition*, pp. 54-61, 2000.

[12] Ming-Hsuan Yang, Narendra Ahuja, "Detecting human faces in Color Images," ICIP, volume 1, pp. 127-130, 1998.

[13] Benjamin D. Zarit, Boaz J. Super, Francis K. H. Quek, "Comparison of Five Color Models in Skin Pixel Classification", *International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-time Systems*, pp. 58-63, September 1999

[14] http://academic.mu.edu/phys/matthysd/web226/L0221.htm